UNIVERSIDAD DE PIURA

TRABAJO FINAL

ANÁLISIS DE DATOS CON PYTHON 2

VIDA UNIVERSITARIA

ANÁLISIS DE MACHINE LEARNING CON DATASET SOBRE LA INDUSTRIA CINEMATOGRÁFICA

ALUMNO:

Nuñez Saravia, Fernando Martín

Universidad de Piura - 06 de febrero del 2022

INTRODUCCIÓN

El presente trabajo utiliza el dataset denominado “Hollywood Theatrical Market Synopsis 1995 to 2021” que contiene los datos de análisis del mercado cinematrográfico basado en el sistema de categorización único “The Numbers”. Específicamente, muestra las películas más vendidas por cada año entre dicho periodo. La data utiliza seis criterios diferentes para identificar las películas según los siguientes atributos: YEAR (año en el que se publicó la película), MOVIE (nombre de la película), GENRE (género de la película en cuestión), MPAA RATING (rating según el “Motion Picture Association of America”) y DISTRIBUTOR (compañía que produjo el filme). Todas las clasificaciones se basan en la venta de entradas (TOTAL FOR YEAR y TOTAL IN 2019 DOLLARS), que se calculan al multiplicar los TICKETS SOLD con algunos de los precios promedio de las entradas anunciados por la MPAA en su informe anual sobre el estado de la industria.

Dicho conjunto de datos contiene varios archivos que ilustran estadísticas como la venta anual de boletos de las películas más taquilleras de cada año desde 1995, así como el género al que corresponden y la empresa que las produjo. Por ello, se diseñarán algunos modelos de regresión en Machine Learning empleando cuatro métodos distintos. Estos métodos serán el Linear Method, K Nearest Neighbors, Random Forest y Stochastic Gradient Boosting. Asimismo, las variables que se analizarán serán los tickets vendidos en todo el tiempo y el total de ventas hasta el 2019, lo cual se intentará predecir las ventas del año en la que fue estrenada. Además de analizar dichas características, el trabajo presenta al inicio del bloc un breve análisis estadístico y las trazas de algunas gráficas que he considerado relevantes.

ANÁLISIS DEL PROBLEMA

Para elaborar dicho sistema, debemos comenzar importando las librerías necesarias. Entre ellas encontramos a Numpy, Pandas, Plotly, Matplotlib y, la más importante para llevar a cabo todo el procedimiento de Machine Learning, Sklearn. Al mismo tiempo, se debe establecer el dataset con el lector CSV que nos ofrece la librería Pandas.

Antes de realizar el análisis estadístico y las gráficas analíticas, debemos preprocesar algunas de las columnas de nuestro dataset. Estas son las tres últimas que, aparentemente son variables numéricas, pero en realidad son variables categóricas en vista de que tienen símbolos como comas y el índice del dólar. Para ello declararemos el comando lambda que nos permitirá eliminar el primer dígito de cada feature (en este caso serán los $) y reemplazar cada una de las comas por un espacio en blanco (por lo que se eliminarán automáticamente). Adicionalmente, se renombraron las etiquetas de las columnas mencionadas para que sigan guardando relación con sus valores.

Posteriormente, se realizará un análisis estadístico relativamente simple. En el caso de las variables numéricas se aplicó el comando ‘describe’ y para las variables categóricas el comando ‘value_counts’; ambas pertenecientes a la librería Pandas.

También se elaboraron un par de gráficas con las librerías Matplotlib y Plotly. En primer lugar, se graficó un histograma con la variable de Géneros para evaluar cuáles fueron aquellos géneros más recurrentes de las películas top de cada año a partir de 1995. En segundo lugar, se estableció una dispersión de datos en tres dimensiones; en los ejes principales estaba el “Año (X)” y los “Tiques Vendidos (y)”, mientras que el color de cada símbolo representaba la variable “Distribuidor” para que a la vez podamos comparar cuál fue la compañía que produjo la mayor cantidad de filmes top en el periodo establecido (por lo visto, Walt Disney abunda demasiado).

Finalmente, se llevaron a cabo el procedimiento de Machine Learning con los cuatro distintos métodos mencionados previamente. Se tuvo en cuenta el siguiente procedimiento: Formar la matriz ‘X’ y el vector ‘y’ → Analizar la dimensionalidad vectorial → Dividir la data → Crear el modelo → Entrenar la data → Validar resultados

Algo a tener en cuenta fue que previo a la división de la data en valores de entrenamiento y de validación, se realizó un escalamiento de la matriz y el vector debido a que eran valores sumamente grandes (millones); y porque los valores de la variable de Tiques Vendidos contenían una cifra menos que las otras dos variables. Por lo tanto, para mejorar los resultados de cada método, se escaló la data.

ANÁLISIS DEL RESULTADO

A continuación, se muestran los coeficientes de correlación, tanto de entrenamiento como de validación, de cada uno de los métodos empleados en el Machine Learning.

```
R2 de entrenamiento con Linear Method: 0.5608589771194735
R2 de validación con Linear Method: 0.6082206835150062

R2 de entrenamiento con KNN: 0.6346498977040056
R2 de validación con KNN: 0.5085897446382659

R2 de entrenamiento con RF: 0.8774248159564757
R2 de validación con RF: 0.458569116294143 76

R2 de entrenamiento con SGB: 0.5059541637491276
R2 de validación con SGB: 0.4959271761221369
```

Tras haber obtenido todos los valores del coeficiente de correlación con cada uno de los cuatro métodos, podemos establecer cierto análisis y arribar las siguientes ideas:

- Por lo visto todos los valores de r2 son ciertamente estables pues oscilan entre 0.4 y 1.0 (lo cual está bastante bien). Ello en parte se debe al escalamiento de la data que se realizó previamente a la creación de los modelos.
- Aparentemente, el valor de r2 de entrenamiento con Random Forest es el mejor entre todos los demás; no obstante, no parece que esté tan bien pues existe cierto sobreajuste en dicho modelo (el r2 de entrenamiento es mucho más mayor que el r2 de validación)
- Por otro lado, respecto al resto de métodos (Linear Method, KNN y SGB), se puede afirmar que están muy bien generalizados pues sus valores de r2 entrenamiento y validación son algo semejantes. Esto se pudo lograr debido al escalamiento de la data pues, previo a ella, obtenía en más de uno de los métodos casos de sobreajuste o subajuste. He ahí la importancia del preprocesamiento de datos al elaborar sistemas (Data Cruda vs Data Procesada)
- Finalmente, es importante hacer un hincapié en el descenso del gradiente, es decir, el algoritmo de optimización de todos los algoritmos de Machine Learning. Se le menciona porque en ocasiones, al correr por segunda vez el programa, se obtenían valores para r2 levemente distintos a los presentados en este informe; pese a ello, el análisis sigue siendo válido para estos modelos.

Tras haber culminado el presente trabajo, se concluye que todos los métodos empleados en este trabajo son adecuados para realizar una regresión de datos con Machine Learning. Lo cual, a su vez, nos permitirá predecir valores futuros para el conjunto de datos determinado.

Podríamos decir que el método de Random Forest tiene ciertas complicaciones a la hora de elaborar el sistema de regresión; en contraparte, este modelo es más útil a la hora de procesar data mucho más compleja (la presentada no lo es tanto). Por último, este análisis es muy útil pues podrá ser aplicado y adecuado a cualquier otro tipo de dataset sobre el mercado cinematográfico e incluso a cualquier otro tipo de mercado si se quisiera replicar.